

# Using Text Similarity to Detect Social Interactions not Captured by Formal Reply Mechanisms\*

Samuel Barbosa  
Institute of Mathematics and Statistics  
University of São Paulo  
São Paulo, Brazil  
Email: sam@ime.usp.br

Roberto M. Cesar-Jr  
Institute of Mathematics and Statistics  
University of São Paulo  
São Paulo, Brazil  
Email: cesar@ime.usp.br

Dan Cosley  
Department of Information Science  
Cornell University  
Ithaca, NY 14853 USA  
Email: danco@cs.cornell.edu

April 19, 2016

## Abstract

In modeling social interaction online, it is important to understand when people are reacting to each other. Many systems have explicit indicators of replies, such as threading in discussion forums or replies and retweets in Twitter. However, it is likely these explicit indicators capture only part of people's reactions to each other; thus, computational social science approaches that use them to infer relationships or influence are likely to miss the mark. This paper explores the problem of detecting non-explicit responses, presenting a new approach that uses tf-idf similarity between a user's own tweets and recent tweets by people they follow. Based on a month's worth of posting data from 449 ego networks in Twitter, this method demonstrates that it is likely that at least 11% of reactions are not captured by the explicit reply and retweet mechanisms. Further, these uncaptured reactions are not evenly distributed between users: some users, who create replies and retweets without using the official interface mechanisms, are much more responsive to followees than they appear. This suggests that detecting non-explicit responses is an important consideration in mitigating biases and building more accurate models when using these markers to study social interaction and information diffusion.

## 1 Introduction

Studies on social networks often use actions people take on other people's online content as evidence of social interactions for developing their models. In domains including Usenet [1], Wikipedia [2], and Facebook [3], explicit replies are interpreted as evidence of interpersonal interaction and social ties. These explicit reactions are also used in studies of influence online, such as predicting when an item is likely to be forwarded in Twitter (e.g., [4, 5]).

Not all responses, however, are explicitly marked by the system. For instance, a post that is explicitly threaded as a reply to a particular post in a discussion forum might nevertheless address another post or posts. In Twitter, the primary focus of this paper, there are buttons for replying to and retweeting another user's tweet—but users might compose a new tweet that references another recently seen without hitting the reply button. Users might do this for a variety of reasons, from being inspired to write their own post on a topic they see coming up in their feed to using the system in ways not intended by the designer (such as copying and pasting content into a new tweet rather than pressing a retweet button).

Being able to identify these non-obvious, indirect responses might allow researchers to have a more accurate view of social interaction than explicit mechanisms provide. This might also improve overall estimates of users' responsiveness to others, for instance, at the individual level, they might indicate how desirable a user is as a follower: people might wish to have followers who are more likely to redistribute their content. Aggregating responsiveness of a user's followers at the ego network level could support better estimates of an individual's

---

\*A final version of this work was published in the 2015 IEEE 11th International Conference on e-Science (e-Science). It can be found in <http://dx.doi.org/10.1109/eScience.2015.31>.

potential reach or influence [6] based on the responsiveness of their followers. Better responsiveness measures could also improve transmission probabilities in epidemiology-inspired models of diffusion in social networks [7].

This paper assesses the prevalence of non-explicit responses in a dataset drawn from Twitter, using a measure of normalized textual similarity between a user’s tweets and recent friends’ tweets based on *tf-idf* scores. Comparing this to the explicit responses provided by the system shows that explicit indicators of response (replies and retweets) in Twitter are in fact associated with high normalized similarity scores. Choosing conservative score cutoffs for predicting that a tweet is a response and manually inspecting high-scoring tweets that are not marked as responses suggests that explicit indicators miss at least 11% of reactions. Further, this varies between users: some users systematically fail to use formal response mechanisms, meaning that these users are under-represented in studies that rely on explicit indicators of response and under-counted when considering their potential as information spreaders. These results show that the problem of non-explicit responses is an important one with practical implications for understanding interaction and influence online.

Such studies often focus on computational models for predicting retweet behavior. For instance, Suh et al. [4] apply Principal Component Analysis to decompose tweets into a space of characteristics, showing that URLs, hashtags, the number of followers and followees, and the age of the account are correlated with retweet behavior. Comarella et al. [5] also find that previous responses to the same tweeter, the tweeter’s sending rate, and the age of a tweet influence retweeting, proposing two ranking methods for reordering tweets to increase retweeting. Petrovic et al. [8] built a *passive-aggressive* classifier for answering that took into consideration social characteristics of the tweets’ author as well as tweets’ textual features, finding that social features are more informative. Peng et al. [9] used *Conditional Random Fields* to model the probability of how a user retweets a message.

Other studies look at variations of the problem. Artzi et al. [10] applied *Multiple Additive Regression-Trees* and *Maximum Entropy Classifiers* to predict both retweets and replies, while Hong et al. [11] model both the binary question of whether a tweet would be retweeted and the eventual number of retweets a message might accrue. Luo et al. [12] and Wang et al. [13] approach a similar problem: given a user and their followers, who will retweet a message generated by the user? Both created classifiers to predict the followers that would retweet a message. Liu et al. [14] studied the social network of questions and answers in *Sina Weibo* looking for characteristics that are associated with a higher number of answers.

These prior works identify a number of useful features that researchers often take into consideration when developing their models. These include textual features of Tweets, user preferences or characteristics, and features of users’ networks including pairwise relationships and graph structure. Table 1 presents a number of these features and the papers that have used them in response prediction. This paper’s focus on the prevalence of implicit responses complements these works by identifying tweets that, although not marked as a response, are in fact likely to be real responses. Such tweets would appear as errors or noise to these models; methods for identifying them might improve both these models and our understanding of why these features matter. For instance, account age might turn out to predict retweet behavior mostly because more experienced users are simply more likely to press the retweet button than new users, rather than having a higher innate propensity to retweet.

When trying to identify non-explicit responses, having a model that explains which messages a user is most likely to be interested can be valuable; that is, the problem of understanding these (message, user) relationships is related to the problem of understanding the (message, reaction) relationships. The main stream of research related to modeling user interests in Twitter is the feed personalization problem, defined by Berkovsky et al. [15] as creating mechanisms that promote and optimize exhibition of interesting content (messages or people, for instance) according to each user’s particular preferences and context. In their survey, they break approaches to feed personalization into three main groups: approaches that consider the pairwise relationship between author and consumer of content, approaches that take into consideration the graph structure of the social network, and approaches that deal with textual information from the users.

As with studies of retweet prediction, feed personalization approaches often use indicators of tie strength as proxies for potential interest. Schaal et al. [16] measure pairwise user similarity through *tf* vectors and topic similarity using LDA. Goyal et al. [17] estimate pairwise influence probability based on the user activity (action log). There are a wide variety of such features; Gilbert and Karahalios [3] estimate pairwise tie strength based on Facebook data based on over 70 features in categories including intensity, intimacy, duration, reciprocal services, structural, emotional support, and social distance.

Network structure also plays an important role in feed personalization. Uysal et al. [18] developed a personalized tweet ranking method based on a retweet metric, useful in reordering feeds or distributing items to users more likely to retweet. Paek et al. [19] asked Facebook users about the perceived importance of items in their timeline, developed classifiers to identify important messages and friends, and studied the predictive power of a number of features including likes, number of comments, presence of links and images, textual information, and shared background information. Both the tie strength and network structure approaches rely on explicit interaction as a tool for estimating tie strength; just as with retweet prediction, being able to identify non-explicit responses might improve these models.

Most related to this paper are text-focused approaches. Text is commonly used in feed personalization, by comparing content similarity of Tweets or users to a user’s previous activity. Hannon et al. [20] developed a system for follower recommendation on Twitter based on *tf-idf* similarity between the users’ newsfeeds. Burgess et al. [21] propose a system to automatically select users when creating lists. The method adopts *tf-idf* to compare content users generated, among other measures and evaluates the performance comparing user-made lists with those generated by the system. This work informs ours by providing evidence that *tf-idf*-based methods are useful in understanding attention and interest.

## 2 Reaction Identification

This section presents the definition of the problem and the method used to attack the identification of non-explicit reactions in Twitter.

When users decide to post a message in Twitter, they might be reacting to some content they saw from one of their followees. The first assumption is that the evidence for these reactions are the textual features in a given tweet by user  $u$  and textual features in the set of recent tweets by  $u$ ’s followees. Another assumption is that, if  $u$  tweeted in reaction to a followee’s message, there should be higher text similarity between that tweet and that message. This work focus on text features, rather than user or network characteristics found in prior work, because they have been shown to be useful while simplifying data collection, computation, and modeling.

This leads to this work’s first research question, about whether text similarity has potential for identifying non-explicit responses. Do explicit responses in fact tend to have high text similarity? If so, what fraction of high-scoring tweets are non-explicit? And, even when similarity is lower, when might non-explicit responses be present?

The second research question asked is how these non-explicit responses are distributed among users. Are many users “invisible” because, although they appear to be responsive based on scores, their responses are not explicit? Why are they lost? Are they naive or low-frequency users who do not know better than to retype or cut and paste or restate? And, is this likely to be important in estimating the overall responsiveness of users?

### 2.1 Influence Window

The information Twitter presents to a user is the set of tweets sent by their followees in reverse chronological order. Comarella et al. [5] study how far back in the user feed is a tweet when replied or retweeted. They divided the users into four sets of increasing levels of activity and found that over 80% of replies and 60% of retweets are responses to one of the 50 most recent tweets in a user’s feed. They also present cumulative distributions of these replied and retweeted tweets when varying the position in the feed, and the last 100 tweets in the feed contain more than 80% of the tweets in these distributions. Based on this, a window  $w_i$  for the tweet  $t_i$  is defined as the last  $n = 100$  tweets generated by user’s  $u$  followees  $f_i$  immediately before  $t_i$ , taken in reverse chronological order. Figure 1 illustrates the window.

### 2.2 Textual Features

Each tweet in a user’s feed also carries associated meta-data besides the message itself, such as the author’s profile and user name, tweet creation time, number of times liked, and number of times retweeted. In this analysis, users are modeled as primarily paying attention to the textual content when considering a response; thus, only textual features the user’s feed exposes are considered.

Tweets are first preprocessed using Python’s NLTK package [22] to be lower case, remove stopwords, and apply Snowball stemming, all common practices when using *tf-idf* scoring. Hashtags, usernames, and processed words are then extracted using the regular expressions shown in Table 2. Finally, the tweet author’s username is added as a feature since that is also visible in the feed.

### 2.3 Message Scoring

The text similarity metric used for this task was the *tf-idf* scoring. It is a proven technique for information retrieval commonly employed in analyzing Twitter data. *tf-idf* stands for term frequency and inverse document frequency. This method takes as input a set of documents  $D$ , where each *document* is a set of *terms*, and produces a document-by-term matrix of *tf-idf* scores. These functions can be scaled, but usually the *tf* is not scaled and the *idf* is logarithmically scaled. For a given (*document*, *term*) matrix entry, the *tf* function is the *term* occurrence count in the *document* and the *idf* function is given by Equation 1.

$$idf(D, term) = \log \frac{|D|}{|\{d \in D | term \in d\}|} \quad (1)$$

Table 1: Some characteristics from online social networks that are commonly used to model users' behavior.

Characteristic	Description
URL	Presence of a link in a tweet. [4, 5, 8–10]
Number of hashtags	Number of hashtags in a tweet. [5, 8–10]
Number of mentions	Number of mentions in a tweet. [4, 5, 8–10, 14]
Number of followers	Number of followers of the author. [4, 8, 10–14]
Number of followees	Number of followees of the author. [4, 8, 10–13]
Presence in lists	Number of times that an author has been added to lists. [8, 12]
Verified	If the author has a verified account. [8, 12]
Ratio of followers over followees	Ratio <i>followers/followees</i> or its inverse. [9, 10]
N-grams	Presence of possible n-grams in the text. Usually used together with dimensionality reduction methods. [8, 10]
Number of Stop Words	Number of stop words in the tweet. [10]
Time	Time when the user received the tweet. [10, 14]
Day of week	Day of the week when the user received the tweet. [10]
Time zone	If the author and the receiver of a tweet are in the same time zone. [12]
Wait time	Average time a user takes to reply or retweet a message. [5, 11]
Timeline position	How many messages on average a user receives between receiving and replying (or retweeting) a tweet. [5]
Tweet age	When the tweet being retweeted was originally created. [5, 11]
Previous interaction	If the user has already replied to or retweeted the author in the past. [5, 12, 13]
Author's activity	Absolute number, frequency, or distribution that represents how the author tweets. [4, 5, 8, 9, 11–14]
Followees activity	Absolute number, frequency, or distribution that represents how the followees of the user tweet. [9]
Tweet size	Number of characters of the tweet. [5, 8]
Author's PageRank	PageRank of the author. [11, 13]
Reciprocal links	If the author and the user follow each other. [9, 11, 13]
Reciprocal followers	Number of followers that the author and the user share. [9, 13]
Reciprocal followees	Number of followees that the author and the user share. [9, 13]
Reciprocal mentions	Number of tweets where the author mentions the user or the user mentions the author. [9]
Reciprocal retweets	Number of retweets that the author and the user share. [9]
Clustering coefficients	Clustering coefficients of the network structure. [11]
Previously retweeted message	If and how many times a message has been retweeted by other users in the past. [4, 11]
Author's retweet count	How many messages of the author have been previously retweeted. [9, 11]
Emoticons	If there is an emoticon in the tweet. [14]
Message topic	Topic identification on the message text or topic similarity measures between the author's interests and the message topic. [9, 12–14]
Language	User's profile language. [8, 13]
Favorite	If the tweet has been marked as a favorite by the author. [4, 8]
Response	If the message received is an answer to a previous message. [8]
Account age	Age of the tweet author's account. [4, 13]
Trending topics words	If the tweet has <i>trending topics</i> ' terms. [8]
Reciprocal hashtags	Number of hashtags in common that the author and the user shared in the past. [13]
Reciprocal URLs	Number of URLs in common that the author and the user shared in the past. [13]
Number of lists	Number of lists that an author created. [13]

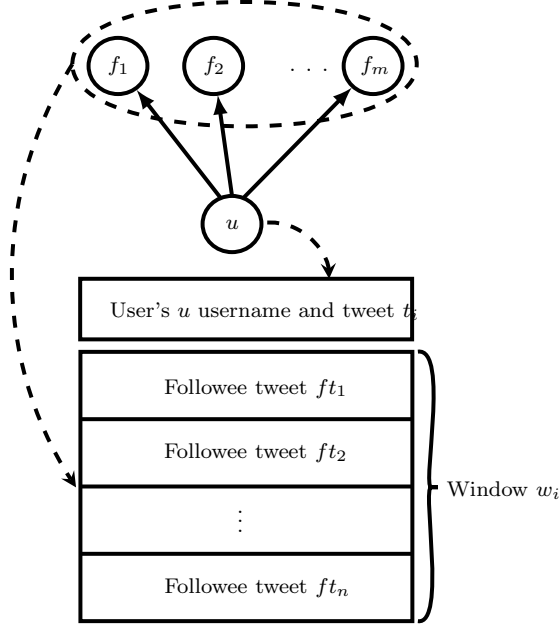


Figure 1: Construction of the window  $w_i$  for the tweet  $t_i$ . The tweets in the window (in this paper,  $n = 100$ ) are those most generated by user  $u$ 's followees most recently before  $t_i$ .

Table 2: Regular expressions used to extract features from tweets.

Hashtags	<code>(?:[\s ^])(#[\w]+)</code>
Users	<code>\B(?:[@])([\w]{1,20})</code>
Words	<code>(?:^[^\s][^\s@#\s\w]*)([\w]+)</code>

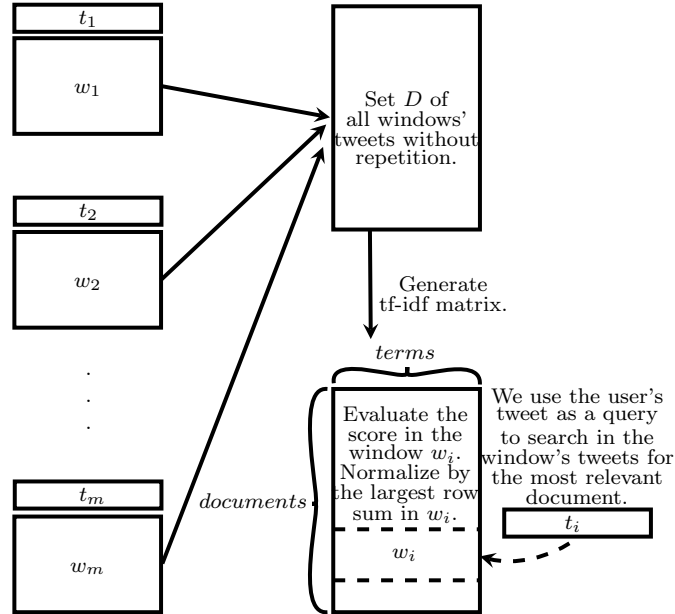


Figure 2: Process to generate each tweet *score*. All the tweets in the windows are used to compose the corpus from which the *tf-idf* matrix for a given user is generated. Each user's tweets are then used as queries to search in their windows for the most relevant followee's tweet.

Notice here that the *idf* is a function of the whole set of documents and a particular *term*, while the *tf* is a function of the document and the *term*. One high level interpretation of these functions is that *tf* indicates how important is *term* for the *document*, while the *idf* captures how common is the *term* among the *documents* and indicates how much information it provides when it occurs in a particular *document*.

The *tf-idf* was calculated using the implementation provided by the Python package scikit-learn [23]. It uses a smoothed version of the *idf* function (even if the *term* happens in all documents it will not be ignored). The

final *tf-idf* document-by-term matrix is given by Equation 2.

$$tf-idf(document, term) = tf * (1 - idf) \quad (2)$$

The set of documents  $D$  is comprised of the tweets in all windows for a user  $u$  (each user has its own set  $D$ , and words in these tweets form a user-specific language model). Each textual feature is one *term* in our analysis, and the *tf-idf* scores matrix is computed for  $D$ .

The tweets generated by  $u$  are then used as queries that leverage the matrix. For each tweet  $t_i$ , its text features are extracted (removing duplicate *terms*) and the *score* evaluated for each pair  $(t_i, ft_j)$ , where  $ft_j$  is a followee’s tweet in  $t_i$ ’s window  $w_i$ . The *pairScore* is given by Equation 3.

$$pairScore(t_i, ft_j) = \sum_{term \in (t_i \cap ft_j)} tf-idf(ft_j, term) \quad (3)$$

To be able to compare in a score-independent way between tweets and users, the score for each tweet is normalized based on the maximum value of the *tf-idf* matrix row sum for the tweets in window  $w_i$ , as given by Equation 4.

$$normalization(w_i) = \max_{t \in w_i} (pairScore(t, t)) \quad (4)$$

This normalization means that the tweet  $t_i$  generated by the user will have a final score of 1 if that tweet reproduces the exact text of the tweet that would yield the maximum score that is present in the window  $w_i$ . The *score* for each tweet  $t_i$  is then given by Equation 5.

$$score(t_i) = \max_{ft_j \in w_i} \frac{pairScore(t_i, ft_j)}{normalization(w_i)} \quad (5)$$

The interpretation of the  $score(t_i)$  is how likely  $t_i$  is to be a response to a friend’s tweet  $ft = argmax_{ft' \in w_i} (pairScore(t_i, ft'))$ .

### 3 Twitter Dataset

This definition of potential response allows ego networks to be collected rather than full network data. This is often a more feasible approach when dealing with online social networks, since even friendly APIs normally impose rate limits. Ego networks are often useful for studying interaction and influence [24, 25]; here, they are appropriate because the method requires only a user’s content and his followees’ in order to reconstruct the feed windows.

The dataset this paper is based on was collected as part of a project to investigate differences in online behavior between political groups, driven by observations that, in the U.S. 2012 presidential election, Democrats were more active and effective in social media than Republicans. This paper draws on that dataset, using ego networks on Twitter belonging to users that followed Barack Obama crawled in the first three weeks of December 2012 using V1.0 of the Twitter API.

The crawler first got all the followers for Obama’s account, then filtered out users that did not choose English as their profile language or had no tweets in the last month. It then randomly selected 547 users and collected up to one month (or the Twitter limit of 3200 historical tweets) of Tweets from each user and all of their followees, creating a set of ego networks.

Because of the 3200 tweet per-user limit, as well as occasional API or network errors, the dataset does not contain a complete record of all followees’ tweets. This could affect estimates of the presence of non-explicit responses; thus, networks where a significant proportion of followees’ tweets appeared to be missing were filtered out. Tweets were considered missing when a followee’s activity only partially overlapped with the ego user’s<sup>1</sup>, with the number of missing tweets estimated based on the length of overlap and the rate of that followee’s tweets. Users for whom over 20% of their followers’ tweets were estimated missing were removed from the dataset, leaving 449 ego networks<sup>2</sup>.

tribution of the time length for the generated windows. Tagged tweets are defined as those indicated by the API as explicit responses, i.e., Replies and Retweets, while the Non-Tagged set is anything not tagged by Twitter<sup>3</sup>.

<sup>1</sup>There is a parallel, opposite problem for users who added followees during the ego user’s activity period; windows for tweets before the followee was added will incorrectly contain their tweets, which the user could not have responded to. We saw no good way to address this and so tolerate the error.

<sup>2</sup>Other thresholds (5%, 10%, 50%, 80%, 100%) were tested. Lower values lead to similar results, while higher values increased the number of users that lacked data for analysis; 20% was chosen as a reasonable trade-off between sample size and meaningfulness of results.

<sup>3</sup>Upper-case names refer to the collected sets in this work, while lower-case names refer to messages in general.

Table 3: Descriptive data. Tweet counts are based on ego users. Each tweet may be tagged either as a retweet or a reply. Replies also provide the replied tweet id, allowing us to count how often a tagged reply refers to a tweet in the window. As with Comarela et al. [5], over 80% of tagged replies reference one of the 100 most recent tweets.

Users	449
Tweets	26051
Average Tweets/User	58.02
Min Tweets/User	1
Max Tweets/User	832
Retweets	5209
Replies	4192
Replies in windows	3455
Window avg. size (h)	5.24
Windows std. deviation (h)	63.87
Windows min size (h)	0.01

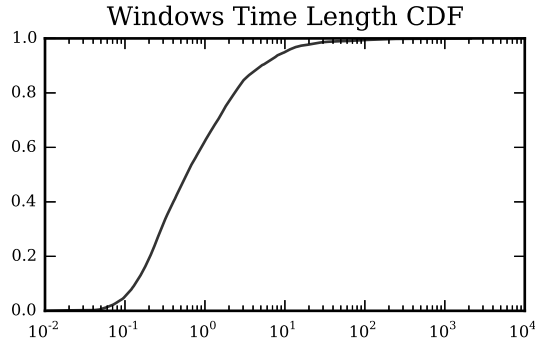


Figure 3: Cumulative distribution function for the time length of the windows given in hours. Most windows’ lengths are in the interval  $[10^{-1}, 10^1]$  hours; about 60% of windows are 1 hour or less, meaning users receive on average over 100 tweets an hour.

Table 4: Sample mean and standard deviation for the normalized similarity score for the Replies, Retweets, and Non-Tagged sets.

	Mean	Median	Std.
Non-Tagged	0.135	0.102	0.136
Replies	0.212	0.200	0.092
Retweets	0.384	0.287	0.282

### 3.1 How prevalent are non-explicit responses?

This section addresses the first research question of whether or not text similarity has potential for identifying untagged responses, starting with whether Tagged reactions indeed tend to have higher scores than Non-Tagged ones.

Mean and median scores are lowest for Non-Tagged and highest for Retweets, as shown in Table 4. This can also be seen in the scores’ histogram for each of these sets in Figure 4. The score behaves as expected when we consider the averages, returning higher values for Replies and Retweets. However, the proximity of the means for the Replies and Non-Tagged and the higher variance of the Non-Tagged makes these two distributions not so well distinguishable based on score alone. The Retweets, on the other hand, present a heavier tail on the distribution. This suggests that the score captures general trends of the Tagged tweets, but is more suitable for Retweets. Considering that the Retweet average is 0.384 and that it is higher than the Replies mean by more than one standard deviation, **high scored messages** are defined as messages with  $score \geq 0.384$ .

Although the Non-Tagged set has a lower average, it has a higher variance than replies. This comes from the fact that Non-Tagged tweets have a heavier tail when compared to replies, as seen in Figure 4. Also, the Non-Tagged high scored tweets are not neglectable when compared with the number of high scored Tagged tweets, as seen in Table 5: such Non-Tagged tweets would comprise about 11% of responses, even with a fairly conservative cutoff of 0.384. However, high scored messages misses most of the explicit Replies with this cutoff choice.

Considering the retweet behavior, it would be expected that the normalized similarity score for retweeted messages would be high as long as the original tweet showed up in the windows and the retweet is basically reproducing the message with almost no modifications. Surprisingly, this is not what is observed in Figure 4. Instead, more than 54% of Retweets have a  $score < 0.384$ . One possible explanation for this is that people

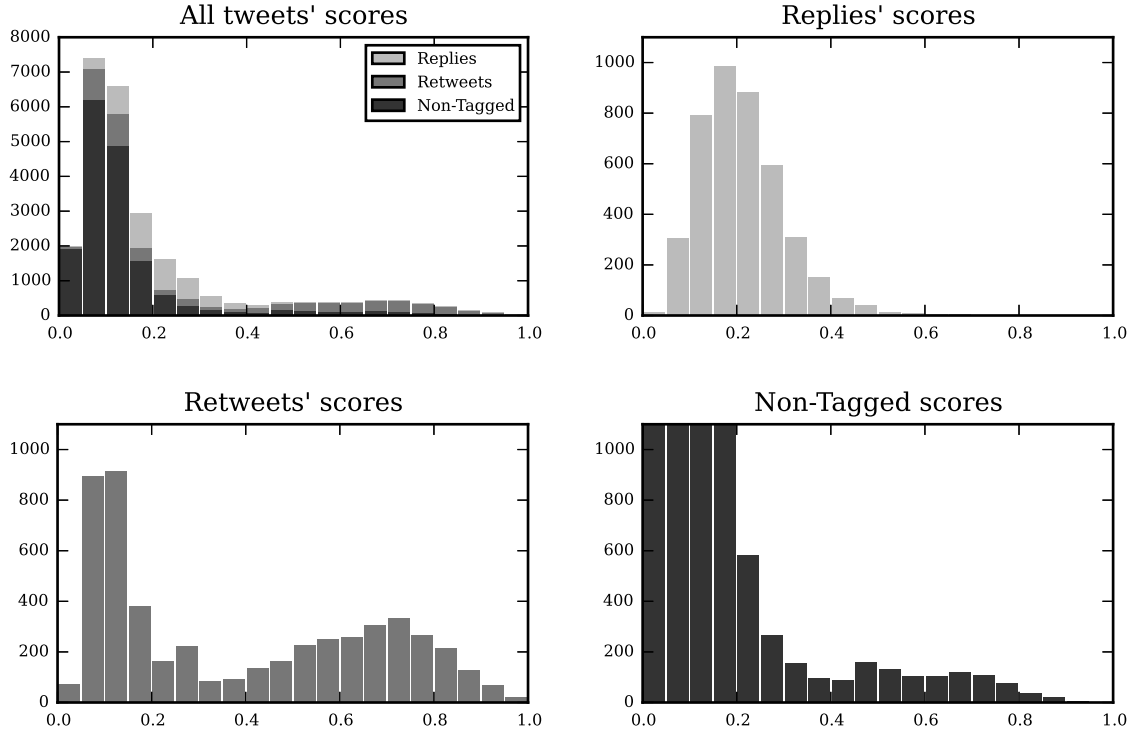


Figure 4: Histograms for the normalized similarity scores. Note that the y-axis for the Non-Tagged subgraph was truncated at 1100 for better visualization of the tail of the distribution and matching other scales. Retweets have a higher average score than Replies, which in turn are higher than Non-Tagged. Further, Retweets have a bimodal distribution; high scores are near-duplicates of the tweets they are responding to, but over 54% have a score below 0.384, suggesting that people often substantially edit retweets or retweet items not in their feed windows.



Table 5: Number of high scored messages and the total of messages for the sets Non-Tagged, Replies and Retweets. The highlighted number of high scored Non-Tagged messages is around 11% of the highlighted total of Tagged messages.

	Non-Tagged	Replies	Retweets
High Scored ( $score \geq 0.384$ )	998	177	2408
Total	16650	4192	5209

sometimes retweet when they use other parts of the interface, such as other users’ profiles or search results, or use social media share buttons attached to tweets on other sites. Another possibility is that people might frequently edit retweets.

### 3.2 Features of Replies, Retweets and Non-Tagged messages

To help understand the mystery of low-scoring retweets, and more generally to understand what sorts of markers the method is using to identify potential responses, a sample of representative tweets from each category across a range of normalized similarity scores is examined. Table 6 (see the Appendix) shows both the user’s tweet (top in each pair) and the text of the highest-scoring followee’s tweet in the window for that tweet (bottom in each pair).

For system tagged Retweets, most of the high scored content has almost the same content as the original message (as expected), as in tweets #1 and #2 in the table. One interesting thing to notice here is that as the tweet length decreases, the normalized similarity score goes down (compare #6 to #1). This is related to the fact that the *tf-idf* score is sensitive to the number of matched words between the query and the document. Below a threshold of around 0.3 in this dataset, this effect disappears. Instead, the text starts to look more like two tweets about a common external topic (#7, #8, #9)—despite the fact that the tweet text preserves the “RT” retweet marker. These would be likely candidates for actual retweets that occur outside the window, either farther back in the feed or other parts of the interface than the feed.

When looking at system tagged Replies, high-scoring replies show two main patterns. In one, they look largely like retweets that were tagged as replies, likely because people pressed the reply button and pasted text from the text they replied to, as in #11. In the other, the tweet mentions multiple users who are conducting an ongoing conversation and want all of them to be notified when someone posts something new, as in #12 and #13. It is important to notice that this set of tweets has a maximum score lower than the other sets; scores on the higher end of the distribution could not be found. Also, it appears that @-mentions are the main source of evidence for the normalized similarity scoring even as it goes down, and in fact, replies with low scores still often look like replies despite the low *tf-idf*. This is often (#16, #19) but not always (#18, #20) indicated by bi-directional @-mentions of the conversational partner.

When looking at Non-Tagged tweets, one of the first things to notice is that high scored tweets usually are retweets that were not captured by the system. In some cases it is likely users are manually copying the content of the messages and adding retweet markers (#23, #24); in others, it is more likely that both users are independently retweeting external content (#21, #22). Users often make small comments together with the original text (#22, #23, #25). As the normalized similarity score goes down, the messages look less like a retweet, but often still appear to be topically related, sometimes via hashtags (#28, #29).

In general, higher normalized similarity scores seem to capture retweets reasonably well, even though being sensitive to their length, and a particular type of reply that involves conversations. Non-tagged tweets with high scores are often retweets or quotes with extra comments from the users, although sometimes the retweets may be common retweeting of external content rather than retweets from the window. Further, even the conservative estimate chosen shows that non-explicit responses are quite common—and it is likely that a number of the of the “middle scoring” tweets are actual responses. Distinguishing those from external influences or underlying interest similarity would be an important next problem in building better models of non-explicit response.

### 3.3 Variations in User Responsiveness

The previous sections demonstrate that it’s likely that 11% or more of Non-Tagged tweets are responses that are not explicitly captured by the system. This section addresses the other main research question of how these losses are distributed among different users in the network.

These Non-Tagged high-scored messages were authored by 129 of the 449 users (29%). This suggests that users generate responses that are missed in a non-uniform way: many users behave as the system expects, using explicit reply and retweet mechanisms, but a significant number respond, at least sometimes, without using those mechanisms. Figure 5 shows histograms for example users that have most or all of their responses untagged by

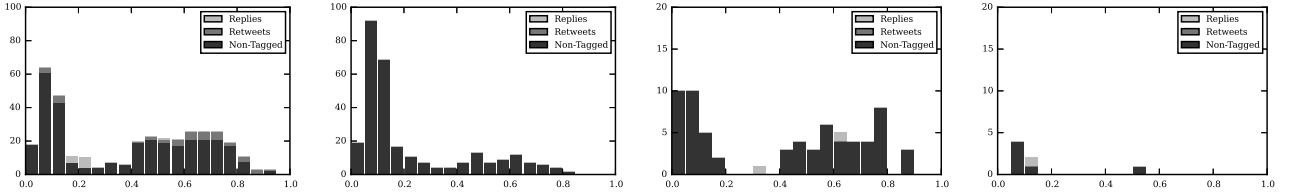


Figure 5: Score histograms for sample users who present a significant amount of high scored Non-Tagged content relative to their total amount of messages, which indicates that most of their reactions are not being properly tagged by Twitter.

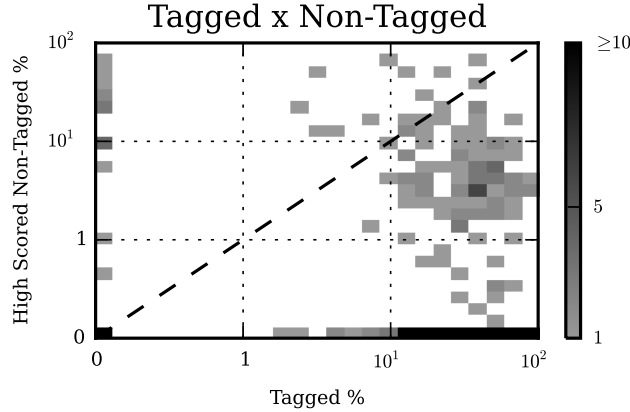


Figure 6: 2D histogram of the percentage of Tagged and high-scored Non-Tagged messages for all users. The scale is linear in the interval  $[0, 1]$  and logarithmic on the interval  $(1, 100]$ ; the dashed line represents an equal percentage of Tagged and Non-Tagged tweets. Many users are non-responsive (the point at the origin) or use the explicit response mechanisms consistently (points hugging the x-axis with a 0 value for high scored Non-Tagged %). However, a significant number never use the explicit response mechanisms (points hugging the y-axis with a 0 value for Tagged %), use them only occasionally (points above the dashed line), or occasionally forget to use them (points below the dashed line).

Twitter even though they present a high *score*. Note that these users span a range of activity levels, meaning that they are not just newbies that don't know how to use the interface.

In order to better understand the behavior distribution among all users, Figure 6 shows a 2d-histogram for the points  $(p_i^T, p_i^N)$ , where each of these points is the percentage of the Tagged messages  $p_i^T$  and the percentage of the high scored Non-Tagged messages  $p_i^N$ . Each of these points is evaluated for a user  $u_i$  in relation to the total number of messages the user authored. The high scored Non-Tagged percentage  $p_i^N$  is the proportion of this user behavior that were likely to be reactions while the percentage  $p_i^T$  is the proportion of reactions actually captured.

The 111 users that never have messages that scored higher than 0.384 nor used explicit system reply mechanisms are concentrated at the origin of the histogram. Users that lay on the  $x$ -axis only react through explicit reaction mechanisms the system offers, therefore have all their reaction Tagged. Similarly, users on the  $y$ -axis never use explicit reaction mechanisms, although they present high scored Non-Tagged content. Users above the dashed line have more high scored Non-Tagged content than Tagged content. It is possible to say that users that lay above the dashed line are more likely to produce content that can be missed by Twitter's tagging system, and they account for 27 users, about 6% of the dataset.

When considering the cumulative distribution of the users according to the percentage of high scored Non-Tagged messages  $p_i^N$ , shown in Figure 7, we identify more than 8% of the users with at least 10% of their messages being high scored and missed by Twitter's tagging system.

These results indicate that methods that rely on explicit indicators of response likely miss or seriously under-represent the behavior of a sizable proportion of the Twitter population.

## 4 Conclusion and Future Work

This paper presented a novel method of capturing some of a user's non-explicit reactions to followees' content in Twitter by using text similarity scores between a user's tweets and those of their followees. The analysis indicates

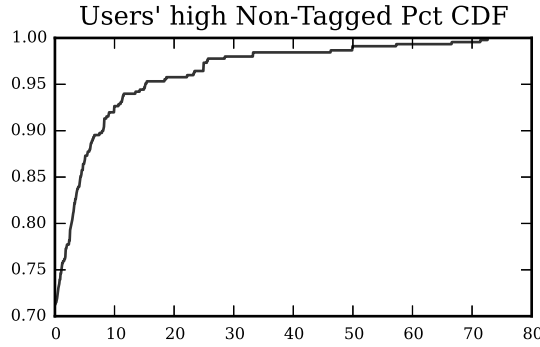


Figure 7: Cumulative distribution of the users for the percentage of high scored Non-Tagged messages. 71% of the users have no high scored Non-Tagged messages, while 8% of the users had at least 10% of their messages high scored and Non-Tagged.

that the method does generate higher scores on average for system tagged Replies and Retweets than Non-Tagged tweets, suggesting that it captures real signal about responses. Using a conservative cutoff for predicting whether a non-tagged tweet is a response suggests that at least 11% of actual responses are not tagged by the system. These responses are distributed across almost a quarter of the users in the dataset, with a quarter of those having more missed reaction messages than explicit system tagged ones. These are not just naive, low-activity users who do not understand Twitter and might be ignored in analysis; a number of these users are quite active, with dozens or hundreds of tweets in a 14-day window.

Although the method has provided useful insights into the prevalence of non-explicit replies in Twitter, it is a coarse model. It tends to under-evaluate Replies; is more sensitive to Retweet size than desirable; likely misses a number of non-explicit responses that have lower scores but are nonetheless real responses to the feed; and doesn't address responses to content outside the feed such as views by hashtag or username. Ongoing work aims at addressing these limitations by improving the quality of the scoring function. One natural way of improving the scoring function is to incorporate other relevant social features highlighted by past work (Table 1). We expect that better models of language, network characteristics, and attention that build on these features would give better estimates of how people react to content produced by their followees.

Another possible unfolding research topic is how to use these reaction scores to understand the reaction patterns and estimate the individual reaction level for each user. This is important for effective models of diffusion at all levels, from understanding when adding an individual to a follower network might be most valuable, to estimating the overall reach of an individual's network, to modeling diffusion of information in the large. Missing 11% of responses and 6% users is a substantial amount of error to bear for such models, making the identification of non-explicit responses an important problem to pursue.

## Acknowledgment

The authors are grateful to FAPESP grant #2011/50761-2, CAPES grant #99999.009323/2014-07, NAP eScience - PRP - USP, NSF grant 1422484, Amit Sharma for his insightful comments, and RepNerv foundation for ongoing support.

## References

- [1] E. Joyce and R. E. Kraut, "Predicting continued participation in newsgroups," *Journal of Computer-Mediated Communication*, vol. 11, pp. 723–747, 2006.
- [2] L. W. Black, H. T. Welser, D. Cosley, and J. M. DeGroot, "Self-Governance Through Group Discussion in Wikipedia: Measuring Deliberation in Online Groups," *Small Group Research*, vol. 42, pp. 595–634, 2011.
- [3] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," *ACM Conference on Human Factors in Computing Systems*, pp. 211–220, 2009. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1518701.1518736>
- [4] B. Suh, L. Hong, P. Pirolli, and E. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," *2010 IEEE Second International Conference on Social Computing (SocialCom)*, pp. 177–184, Aug. 2010. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5590452>[http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=5590452](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5590452)

- [5] G. Comarella and M. Crovella, “Understanding factors that affect response rates in twitter,” *HT '12 Proceedings of the 23rd ACM conference on Hypertext and social media*, 2012. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2310017>
- [6] P. Domingos and M. Richardson, “Mining the Network Value of Customers,” *Proceedings of the Seventh {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining*, pp. 57–66, 2001. [Online]. Available: <http://doi.acm.org/10.1145/502512.502525>
- [7] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, “The role of social networks in information diffusion,” *WWW 2012 Session: Information Diffusion in Social Networks April 1620, 2012, Lyon, France*, pp. 519–528, 2012. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2187907><http://arxiv.org/pdf/1201.4145>
- [8] S. Petrovic, M. Osborne, and V. Lavrenko, “RT to Win! Predicting Message Propagation in Twitter.” *ICWSM '11 International AAAI Conference on Weblogs and Social Media*, 2011. [Online]. Available: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2754/3209>
- [9] H. Peng, J. Zhu, and D. Piao, “Retweet modeling using conditional random fields,” *ICDMW '11: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 336–343, Dec. 2011. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6137399>[http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=6137399](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6137399)
- [10] Y. Artzi, P. Pantel, and M. Gamon, “Predicting responses to microblog posts,” *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2382126>
- [11] L. Hong, O. Dan, and B. Davison, “Predicting popular messages in twitter,” *WWW '11 Proceedings of the 20th international conference companion on World wide web*, p. 57, 2011. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1963192.1963222><http://dl.acm.org/citation.cfm?id=1963222>
- [12] Z. Luo, M. Osborne, J. Tang, and T. Wang, “Who Will Retweet Me? Finding Retweeters in Twitter,” in *Proceedings of the 19th International Conference on World Wide Web*, 2013, pp. 5–8. [Online]. Available: <http://homepages.inf.ed.ac.uk/miles/papers/sigir13a.pdf>
- [13] X. Wang, H. Liu, P. Zhang, and B. Li, “Identifying Information Spreaders in Twitter Follower Networks,” School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tech. Rep., 2012. [Online]. Available: <http://dmml.asu.edu/users/xufei/Papers/TR-12-001.pdf>
- [14] Z. Liu and B. Jansen, “Factors influencing the response rate in social question and answering behavior,” *CSCW '13 Proceedings of the 2013 conference on Computer supported cooperative work*, p. 1263, 2013. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=2441776.2441918><http://dl.acm.org/citation.cfm?id=2441918>
- [15] S. Berkovsky and J. Freyne, “Personalised Network Activity Feeds: Finding Needles in the Haystacks,” in *Mining, Modeling, and Recommending 'Things' in Social Media*, 2015, vol. 8940, pp. 21–34. [Online]. Available: <http://link.springer.com/10.1007/978-3-319-14723-9>
- [16] M. Schaal, J. O'Donovan, and B. Smyth, “An analysis of topical proximity in the twitter social graph,” *Social Informatics*, pp. 232–245, 2012. [Online]. Available: [http://link.springer.com/chapter/10.1007/978-3-642-35386-4\\_18](http://link.springer.com/chapter/10.1007/978-3-642-35386-4_18)
- [17] A. Goyal, F. Bonchi, and L. V. Lakshmanan, “Learning influence probabilities in social networks,” *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, p. 241, 2010. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1718487.1718518>
- [18] I. Uysal and W. B. Croft, “User oriented tweet ranking: A filtering approach to microblogs,” *International Conference on Information and Knowledge Management, Proceedings*, pp. 2261–2264, 2011. [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-83055179234&partnerID=40&md5=9a204d1a80c65f31e0b894c55c9a4737>
- [19] T. Paek, M. Gamon, S. Counts, D. M. Chickering, and A. Dhesi, “Predicting the Importance of Newsfeed Posts and Social Network Friends,” *Artificial Intelligence*, pp. 1419–1424, 2010.
- [20] J. Hannon, K. McCarthy, and B. Smyth, “Finding Useful Users on Twitter: Twittomender the Followee Recommender,” *Springer-Verlag Berlin Heidelberg*, vol. 6611, pp. 784–787, 2011. [Online]. Available: <http://www.springerlink.com/index/6817783481218777.pdf>

- [21] M. Burgess, A. Mazzia, E. Adar, and M. Cafarella, “Leveraging Noisy Lists for Social Feed Ranking,” *Association for the Advancement of Artificial Intelligence*, 2013. [Online]. Available: <https://cond.org/noisylists.pdf>
- [22] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. O’Reilly Media, Inc., 2009.
- [23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [24] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith, “Finding social roles in Wikipedia,” *Proceedings of the 2011 iConference*, pp. 1–11, 2011. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1940778>
- [25] A. Sharma, M. Gemici, and D. Cosley, “Friends, Strangers, and the Value of Ego Networks for Recommendation.” *ICWSM*, 2013. [Online]. Available: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6115/6333>

# A Appendix

Table 6: Pairs of users' tweets (top in each row) and highest scoring messages in the windows (bottom in each row) for Retweets, Replies, and Non-Tagged tweets. Tweets were randomly selected across the range of scores in each set.

#	Score	Retweets
1	1.0	brandonlondon: RT @neiltyson: A H R B Q D W E F L M N S X G I J K O P C T V Y U Z -- Gotta love what the alphabet looks like in alphabetical order. neiltyson: A H R B Q D W E F L M N S X G I J K O P C T V Y U Z -- Gotta love what the alphabet looks like in alphabetical order.
2	0.834	michael.palko: RT @8.Semesters: A girlfriend would be great, but I'm already in a pretty committed relationship with alcoholism and bad decisions. 8.Semesters: A girlfriend would be great, but I'm already in a pretty committed relationship with alcoholism and bad decisions.
3	0.768	mike.sprague: RT @mshowalter: All the weird horny stuff between Glenn and Maggie on Walking Dead makes me very uncomfortable. mshowalter: All the weird horny stuff between Glenn and Maggie on Walking Dead makes me very uncomfortable.
4	0.602	_ShesBrownSKIN: RT @CarGotThat: Brandywine Came Out With Win, #TeamBwine _CarGotThat: Brandywine Came Out With Win, #TeamBwine
5	0.522	Becchappell: RT @olivaaaajayne.: I just love Toy Story, all of them olivaaaajayne.: I just love Toy Story, all of them
6	0.408	_tiki: RT @Greektown1921: Welp now you know Greektown1921: Welp now you know
7	0.303	terrigolas: RT @jam_bu88: Facebook is down? Oh no, how are cancer and child abuse going to stop without all those likes? :( dsilverman: RT @CalebGarling: Stop acting like we have rights' on Facebook http://t.co/geE9NjHH
8	0.248	Becchappell: RT @j4kebro: going to be slightly awkward when Jahmene scans the xfactor winner's single in Asda bombaytricycle: RT @justaholyfoool: Jahmene is gonna be scanning James Arthers CD at ASDA now.. awks
9	0.132	HollywoodLadyj: RT @RealSkipBayliss: RG3 should give Michael Vick a class in scrambling. ESPN_FirstTake: RG3 running at 4G!
10	0.070	davidAmejia: RT @Snoopy: It's Monday, Snoopy! http://t.co/as0F9yPA AshKetchum151: Mondays are like Zubats. Nobody likes Zubats.
#	Score	Replies
11	0.768	esterrick: RT @StationBistro1: On deck - Next week's soup is White Bean and Smoked Turkey Chili! StationBistro1: On deck - Next week's soup is White Bean and Smoked Turkey Chili!
12	0.693	Serrae: @MollytheGhost @PhantomRat @hollye83 @hockeybychoice @onlymystory @sjopierce @phouse1964 Hate them. hollye83: @hockeybychoice @onlymystory @PhantomRat @sjopierce @phouse1964 @MollytheGhost @Serrae Hateful. Just hateful.
13	0.573	HectorBesmonte: @EmmittWard @jccassiel @hottiemarkie33 @mark_purdie @Rhino108 @JasonReedyOH420 yeah! thanks emmit! muah! love, Hugs! for you! EmmittWard: @HectorBesmonte @jccassiel @hottiemarkie33 @mark_purdie @Rhino108 @JasonReedyOH420 happy birthday hector sending you love and hugs buddy.
14	0.477	VinnyG5: @shanmilanowski I wasn't really that drunk that day...I wouldn't get hammered and let you drive with me..but that's a secret so shhhh shanmilanowski: RT @VinnyG5: @shanmilanowski lets do that thing were we get drunk and drive around the city while I'm playing my guitar .....again.
15	0.386	RazWorth: @SophieRaby yeah! Buzzin SophieRaby: @RazWorth do you? :(
16	0.283	SarahMcCallumXX: @SarahMcCallumXX @mton1996 forgot the x hahaa mton1996: @SarahMcCallumXX aw hen, I feel for you x
17	0.245	missRaichl: @michel.andness good morning! michel.andness: Good morning everyone.
18	0.168	Mahalia.Enares: @kiafranklins. HAPPY BIRTHDAY!:) xx istoleursmartie: @Mahalia.Enares i can be!!
19	0.133	MeganDoesNOLA: @KurlyKonfektion Niieee... KurlyKonfektion: @MeganDoesNOLA lmao! I'm gonna put a slice of bacon with/in my drink and see what happens lol
20	0.068	essfardella: @Ali_Diesel_ There it is. Ali_Diesel_: RT @shkeeberr: I am not a slut. I'm an erection enthusiast.
#	Score	Non-Tagged
21	0.920	hypervocal: RT @Reuters: FLASH: #Egypt's Mursi has left presidential palace, two presidency sources say after protesters, police clash outside. AntDeRosa: RT @Reuters: FLASH: #Egypt's Mursi has left presidential palace, two presidency sources say after protesters, police clash outside.
22	0.884	hypervocal: It's time. RT @whitehouse: hey guys - this is barack. ready to answer your questions on fiscal cliff & #my2k. Let's get started. -bo ethanklapper: RT @whitehouse: hey guys - this is barack. ready to answer your questions on fiscal cliff & #my2k. Let's get started. -bo
23	0.782	Mrjscott: A white woman... RT @T_dot_Lee: A woman? RT @majic1021: Fresh Prince' Star Alfonso Ribeiro Weds http://t.co/lft3Zqlr T_dot_Lee: A woman? RT @majic1021: Fresh Prince' Star Alfonso Ribeiro Weds http://t.co/iTrfZfeM
24	0.697	wulan.kyuuifilan: RT @WestlifeFansite: hear @nickybyrneoffic on the radio one minute ago!! it was funny :D x WestlifeFansite: hear @nickybyrneoffic on the radio one minute ago!! it was funny :D x
25	0.579	esterrick: RIP Mr Brubeck. Take five. "@annesaurus: Dave Brubeck, jazz icon, dead at 91. http://t.co/ae9UIRmP" Supperphilly: RT @annesaurus: Dave Brubeck, jazz icon, dead at 91. http://t.co/sOrB0FBR
26	0.443	Zac.Hartlage14: @BadJerry20 OKC traded James Harden 24_Jag: Why WOULD OKC TRADE JAMES HARDEN????
27	0.359	Serrae: (that should have had a link to the Tina and Amy host Golden Globes article. But I'm too lazy to fix it now) MichaelAusiello: Genius Move: Tina Fey and Amy Poehler to Host 2012 Golden GlobeAwards! http://t.co/zdc2hS8F
28	0.275	nicoleoraha: @Niallofficial are you excited to come to Australia and meet all your amazing fans like me? ;) xx #asknialler 11 ahoynialler: @Niallofficial EXCITED TO COME BACK TO AUSTRALIA, cause we miss you lots xox #asknialler
29	0.242	Serrae: All of @fatherdowling's #captainhottie pirate puns for #ouat are perfect. It's the reason we are twitter friends. fatherdowling: I apologize in advance for inappropriate pirate puns. #sorryImotsorry #OUAT
30	0.139	ESTL63: Forever my lady lol DaMontesMom_415: @VictoriaMathis just go get it!! Lol (the devil) aint I? Lol but I would lol